Title: Two Perspectives on the Generalizability of Lessons from Scaling Up SimCalc

Author(s):

Jeremy Roschelle, SRI International
Deborah Tatar, Virginia Tech
Larry Hedges, Northwestern
Nicole Shechtman, SRI International

Background/context:
At the 2008 SREE conference, we reported the results of two large-scale randomized experiments that addressed the research question "Can a wide variety of teachers use an integration of curriculum, software, and professional development to increase student learning of complex and conceptually difficult mathematics?" In both studies, one targeting 7th-grade mathematics and the other 8th-grade mathematics, the main effect was statistically significant and showed that students using a three-week replacement unit learned more than students in a business-as-usual condition. The student-level effect sizes were 0.63 and 0.50 respectively. These results were robust across the demographic groups and socioeconomic settings that were sampled. This Scaling Up SimCalc research has the potential to influence practice both by identifying a particular genre of software that may be particularly effective – software featuring dynamic mathematical representations – and the importance of integrating software, curriculum, and teacher professional development in order to achieve scalable implementations. However, to have the strongest influence on practice, issues regarding generalizability must be addressed.

Purpose / objective / research question / focus of study:
One purpose of educational research is to provide information about the likely impact of interventions or treatments on policy-relevant populations of students. Randomized experiments are useful for estimating the causal effects of interventions on the students in schools that participate in the experiments. Unfortunately, the samples of schools and students participating in experiments, including ours, are typically not probability (random) samples. Thus, even well-conducted experiments may not yield results that generalize to populations of interest. In the Scaling Up SimCalc experiments, one concern about the sample is that teachers were volunteers and potentially not representative of a broader teaching population. Although the volunteer teachers were randomly assigned to condition (reducing the chance that results were due to selection bias), the properties of the volunteer pool as a whole might limit generalizability to broader or differently-selected populations. A second concern is that, because pragmatic issues unrelated to sampling led to recruitment in regions with high proportions of Hispanic and Caucasian students and teachers, other groups of interest, such as African-American students and teachers, were underrepresented in the studies.

In light of these and other concerns, this paper examines generalizability from two complementary perspectives. First, we have conducted detailed analyses of the characteristics of teachers and schools participating in the sample in comparison to others in the state in which the experiments took place. Second, we present findings from a novel statistical method developed to permit principled generalization from research samples to well-defined populations.

Setting:
The studies took place during the 2005-06 and 2006-07 school years in 115 middle schools throughout several geographic regions across the state of Texas. Texas is an ideal state for scaling research, as it already has an aligned system of standards, curriculum, assessment, and teacher professional development, as well as a widely diverse student population with respect to ethnicity, socioeconomic status, and urbanicity. Targeted geographic regions included the large metropolitan areas of Austin, Dallas, and Fort Worth; smaller cities and more suburban and rural areas of Midland, Wichita Falls, and Lubbock; and the largely Hispanic and socioeconomically disadvantaged Rio Grande region along the Mexican border.

Population / Participants / Subjects:
Teachers were recruited through a statewide network of Educational Service Centers (ESCs) coordinated through the Dana Center at the University of Texas at Austin. The ESCs provide professional development support and training for teachers throughout the state. In order to ensure that the recruitment process was not biased by the types of relationships the ESCs had with particular schools or teachers, the researchers gave each ESC a randomized list of schools in their region and asked the ESCs to contact schools in that order. We believe the ESCs contacted schools in an order that reflected a balance between convenience and the technique we proposed. Nonetheless, schools and teachers were still (paid) volunteers. Thus we would not argue that our samples were random.

Table 1, 2, and 3 show the sample sizes and key characteristics of the teachers, schools and students in the two studies.

(please insert Tables 1, 2, 3 here)

Intervention / Program / Practice:
SimCalc interventions are an integration of three elements – software, curriculum, and professional development. In this systems approach, we do not make claims that any one of these elements is more important than the others. Participating teachers used software and curriculum materials, and received professional development designed specially for this experiment. The design goals were to exemplify the SimCalc approach and meet Texas state standards.

Hallmarks of the SimCalc approach to the mathematics of change and variation are:

1. Anchoring students' efforts to make sense of complex mathematics in their experience of familiar motions, which are portrayed as computer animations.
2. Engaging students in activities in which they make and analyze graphs that control animations.
3. Introducing piecewise linear functions as models of everyday situations with changing rates.
4. Connecting students' mathematical understanding of rate, proportionality, and linear function across key mathematical representations (algebraic expressions, tables, graphs) and familiar representations (narrative stories and animations of motion).
5. Structuring pedagogy around a cycle that asks students to make predictions, compare their predictions to mathematical reality, and explain any differences.
6. Integrating curriculum, software, and teacher professional development as mutually supporting elements of implementation.

Figure 1 shows the key SimCalc MathWorlds software features used in these experiments that allow students to manipulate graphs and algebraic expressions that describe linear and piecewise linear motion.

(please insert Figure 1 here)

Table 4 elaborates the mathematical content covered in the SimCalc interventions. Two curriculum units were designed for these studies. The 7^{th} grade curriculum—Managing the Soccer Team— addresses central concepts of proportionality: linear function in the form $y = kx$,

and rate. The 8th grade curriculum—Designing Cell Phone Games—addresses linear function and average rate. The materials for both units were designed to be used daily over a 2–3 week period, replacing regular lessons on the same topics. Professional development for teaching these units consisted of a sequence of workshops totaling 6 days in each study. The workshops were training and planning opportunities with mathematical content, SimCalc software, and curriculum materials.

(please insert Table 4 here)

Research Design:
Each study was a randomized controlled experiment with pre/post measures. Teachers were randomly assigned to either a Treatment group, which received the SimCalc intervention as outlined above, or a Control group. The counterfactuals were designed such that Control teachers would receive professional development of quality and usefulness similar to the SimCalc intervention, but would not receive the SimCalc intervention and would be asked to teach the parallel content as usual. Random assignment occurred at the school level to avoid contamination between conditions within a school and provide teachers in the same school with a community of practice. Note, however, that most schools had only one participating teacher.

Figure 2 shows the experimental designs and timelines for each study. In the 7th Grade study, both the Treatment and Control groups received the Texas professional development workshop, "TEXTEAMS," which provides important mathematical foundations for understanding proportionality. Control teachers were then asked to teach rate and proportionality as usual in their classrooms. In addition, the 7th Grade study was a delayed treatment design in which Control teachers were promised and provided the complete SimCalc intervention in a second year. In the 8th Grade study, Control teachers were provided a workshop of equal quality and relevance to their teaching (Teaching Mathematics TEKS Through Technology, "TMT3," which focused on the content of statistics) and were asked to teach linear function as usual during the school year. In addition, to examine the effects of fading research team support, we implemented a train-the-trainer professional development model in which workshop leaders were trained to deliver the teacher workshops. This is in contrast to the 7th grade studies in which the curriculum designer delivered the teacher workshops.

(please insert Figure 2 here)

Data Collection and Analysis:
The measures were as follows. The primary dependent measures in both studies were student learning of relevant mathematical concepts (see Table 4 for content). Pretests were administered pre-unit, and posttests were administered post-unit. Key teacher measures included assessments of teacher mathematical knowledge for teaching; a questionnaire about teacher background, attitudes, and beliefs; a teacher log about the target class; a daily log in which teachers gave a structured report of their implementation of the unit; and a teacher retrospective log about the unit as a whole. In addition, demographic data about each participating school was drawn from the Texas Public Education Information Management System (PEIMS) datasets. PEIMS is maintained and distributed by the state of Texas and reports the results of a complete census of teachers, schools, and districts conducted yearly. Teachers also did a one-hour telephone interview about their experiences in the program.

Hierarchical linear modeling (HLM) was employed to estimate the effects of the treatment (Raudenbush & Bryk, 2002). HLM allows accounting for measurement and sampling error at both the student and classroom level, resulting in correctly adjusted standard errors for the treatment effect. While random assignment occurred at the school level, we used two-level models (students nested within classrooms) because most schools had only one teacher.

Our first additional analysis of generalizability compared teachers, classrooms and schools in the samples to two additional populations of teachers, classrooms and schools. As the study was conducted only in certain regions of Texas, we compared the sample group to the population throughout those regions. Further, we compared the sample group to the population in the entire state of Texas. Population data was extracted from the PEIMS database, which includes measures of school poverty, school ethnicity, school size, teacher gender and ethnicity, teacher certification, teaching experience.

The second additional analysis of generalizability estimates the population average treatment effect, as well as the uncertainty (standard error) of that estimate, providing a quantification of the degree of uncertainty in generalizing from the experiment to the population of interest (see Smith, 1983). It involves first stratifying the population and the experiment on the independent factors of interest and then computing the treatment effect in each stratum of the experimental sample. The population average treatment effect is estimated as a weighted mean of the stratum-specific treatment effects. To avoid the difficulty of evaluating separate treatment effects in the very large numbers of strata that occur when even a small number of contextual factors are considered jointly, this method uses propensity scores to summarize the covariates and stratify on propensity scores, which is virtually as effective in matching the population to an experimental sample (Cochran, 1968).

Findings / Results:

As previously reported, in both studies, the main effect was statistically significant and showed that students in the Treatment group learned more than students in the Control group (see Figure 3). As shown in Figure 3, the effect sizes were large overall. In both studies, the difference between the groups occurred mostly on the complex portions of the assessments. The effect sizes of the treatment on the simple portion were small. This may be because the students had high pretest scores on simple mathematics, suggesting they had already learned it.

(please insert Figure 3)

The focus of this paper is not differences between control and treatment groups; these differences were few and minor; we have argued elsewhere (Tatar & Stroter, 2009) that these differences were not a threat to the validity of the overall experimental results. This paper focuses on potential differences between the sample and larger populations.

The first generalizability analysis found that the sample involved a variety of teachers in terms of age, experience level, attitudes and teaching philosophy and a variety of campus locations, school sizes, and ethnicities. With respect to most variables, we could not detect differences in either the means or range between our sample and the regions or between our sample and the

whole state of Texas. This suggests that the results can be generalized to broader populations and settings. However, there were some key factors that limit generalizability in specific ways. First, neither of the experiments included a large urban campus – mostly likely because large urban campuses tend to have their own professional development centers and thus have weaker ties to ESCs. Second, the African-American population was not well represented in any of our studies, either at the teacher or the student level. These findings suggest that further research is necessary to determine whether the SimCalc interventions would be effective in large urban campuses and schools with large African-American populations.

The second generalizability analysis is not complete as of this writing, but will be complete before the SREE conference. Performing this analysis on the SimCalc dataset will both test the new statistical method and provide valuable information on how we should or should not generalize from our findings to larger populations, such as the state of Texas.

Conclusions:

Although it would be desirable for rigorous experiments to use probability or random samples of the target populations, in practice this is nearly impossible to achieve. Recruiting schools is difficult and under most circumstances, it is not possible to achieve a big enough pool of schools such that they can be selected randomly. Likewise, it is often not possible to sufficiently sample all populations of interest within the scope of a particular experiment, due to pragmatic, logistic, and financial limitations. Hence, the process of translating from research to practice is often limited by questions of generalizability from actual samples to broader populations.

Our research in the context of the Scaling Up SimCalc experiments has led us to propose the use of complementary approaches for examining generalizability. A foundational approach is to start out with the best sampling procedure possible, striving to achieve a broad and representative sample. We provided our recruiters with randomized lists of schools, but found that the actual schools they contacted reflected a tension between random selection and convenience. We complemented this procedure with two additional analyses. The first found some ways in which our samples do not reflect the full diversity of Texas; in particular we did not have large urban districts and did not have many African American participants. However, with regard to many other characteristics, our sample is not systematically different from the full population in the state of Texas. Our hypothesis is that the second analysis will predict positive effects for all populations of interest, but with wider confidence intervals for populations that were undersampled. Thus, we should have good confidence in how our results generalize to Hispanic schools but less confidence as to how the generalize to African American schools. Overall, this affects how we share the results of our research with the practitioner community.

The overall contribution of this research is to suggest a strategy for using experiments that are conducted on nonprobability (nonrandom) samples to draw inferences about the average treatment effects in well-defined, policy-relevant populations, taking full account of the clustered structure of the experimental data. The significance to the field is that our approach permits policy researchers to use the same experiment to evaluate the likely effect of an intervention in different policy-relevant populations (e.g., different cities or states). The method quantifies the uncertainty involved in these inferences, revealing the limits of generalizations that are possible.

Appendices

Appendix A. References

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, 24, 295–313.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods, Second edition.* Newbury Park, CA: Sage Publications.

Smith, T. M. F. (1983). On the validity of inferences from non-random samples. *Journal of the Royal Statistical Society*, Series A, 146, 394–403.

Tatar, D., & Stroter, A. (2009). *Recruitment strategies, outcomes, and implications for a randomized controlled experiment with teachers* (No. 3). Menlo Park, CA: SRI International.

Appendix B. Tables and Figures

	7th Grade		8th Grade	
Group	Teachers	Students	Teachers	Students
Control	47	825	23	303
Treatment	48	796	33	522
Total	95	1,621	56	825

Table 1. Sample sizes by study and group

	7th Grade Study		8th Grade Study	
Variable	**Control**	**Treatment**	**Control**	**Treatment**
Total count	47	48	23	33
Percent female	81	77	82.6	84.8
Years teaching total (mean)	10.5 Range: 1 – 29	12.4 Range: 1 – 40	9.6 0 – 27	7.9 0 – 31
Years teaching mathematics (mean)	9.5 Range: 1 – 29	11.0 Range: 1 – 40	9.9 0 – 27	8.2 1 – 32
Teacher ethnicity				
Percent white	70.2	77.1	87.0	78.8
Percent Hispanic	25.5	20.8	8.7	15.1
Percent Asian	4.2	*0*	0	0
Percent African-American	0	2.1	4.3	6.0
Percent with a master's degree	17.0	18.8	26.1	6.0

Table 2. Teacher-level characteristics of the samples.

	7th Grade Study		8th Grade Study	
Variable	**Control**	**Treatment**	**Control**	**Treatment**
Total count of schools	37	36	19	23
Percent free lunch (mean)	53 Range: 3 – 99	54 Range: 2 – 94	43 0 – 89	42 0 – 92
Campus ethnicity (mean)				
Percent white	44	47	61	55
Percent Hispanic	49	45	28	36
Percent Asian	2	2	2	2
Percent African-American	5	5	9	7

Table 3. School-level characteristics of the samples.

	M_1 – Conceptually Simple	M_2 – Conceptually Complex
Overview of Concepts	*Concepts are typically covered in the grade-level standards, curricula, and assessments*	*Building on the foundations of M_1 concepts, constitute more complex building blocks of the mathematics of change and variation found in algebra, calculus, and the sciences*
7th Grade Studies (focus on rate and proportionality)	• Simple a/b = c/d or y = kx problems in which all but one of the values are provided and the last must be calculated • Basic graph and table reading without interpretation (e.g., given a particular value, finding the corresponding value in a graph or table of a relationship)	• Reasoning about a representation (e.g., graph, table or y = kx formula) in which a multiplicative constant "k" represents a constant rate, slope, speed, or scaling factor across three of more pairs of values that are given or implied • Reasoning across two or more representations
8th Grade Study (focus on linear function)	• Categorizing functions as linear/nonlinear and proportional/ nonproportional • Within one representation of one linear function (formula, table, graph, narrative), finding an input or output value • Translating one linear function from one representation to another	• Interpreting two or more functions that represent change over time, including linear functions or segments of piecewise linear functions • Finding the average rate over a single multi-rate piecewise linear function

Table 4. Core mathematical concepts in the studies' curricula and assessments

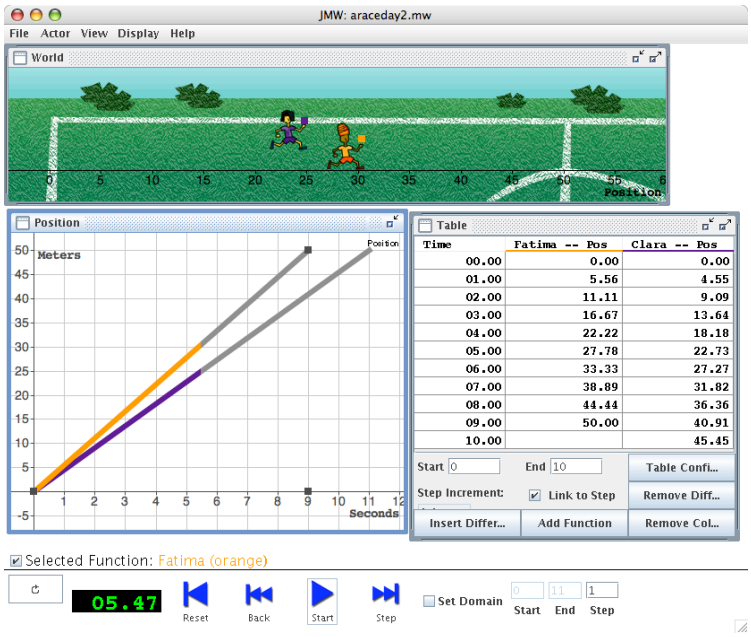


Figure 1. SimCalc MathWorlds software, showing a graph, table, and animation. All mathematical representations are linked so that changes in one representation are immediately reflected in the other representations of the same function.

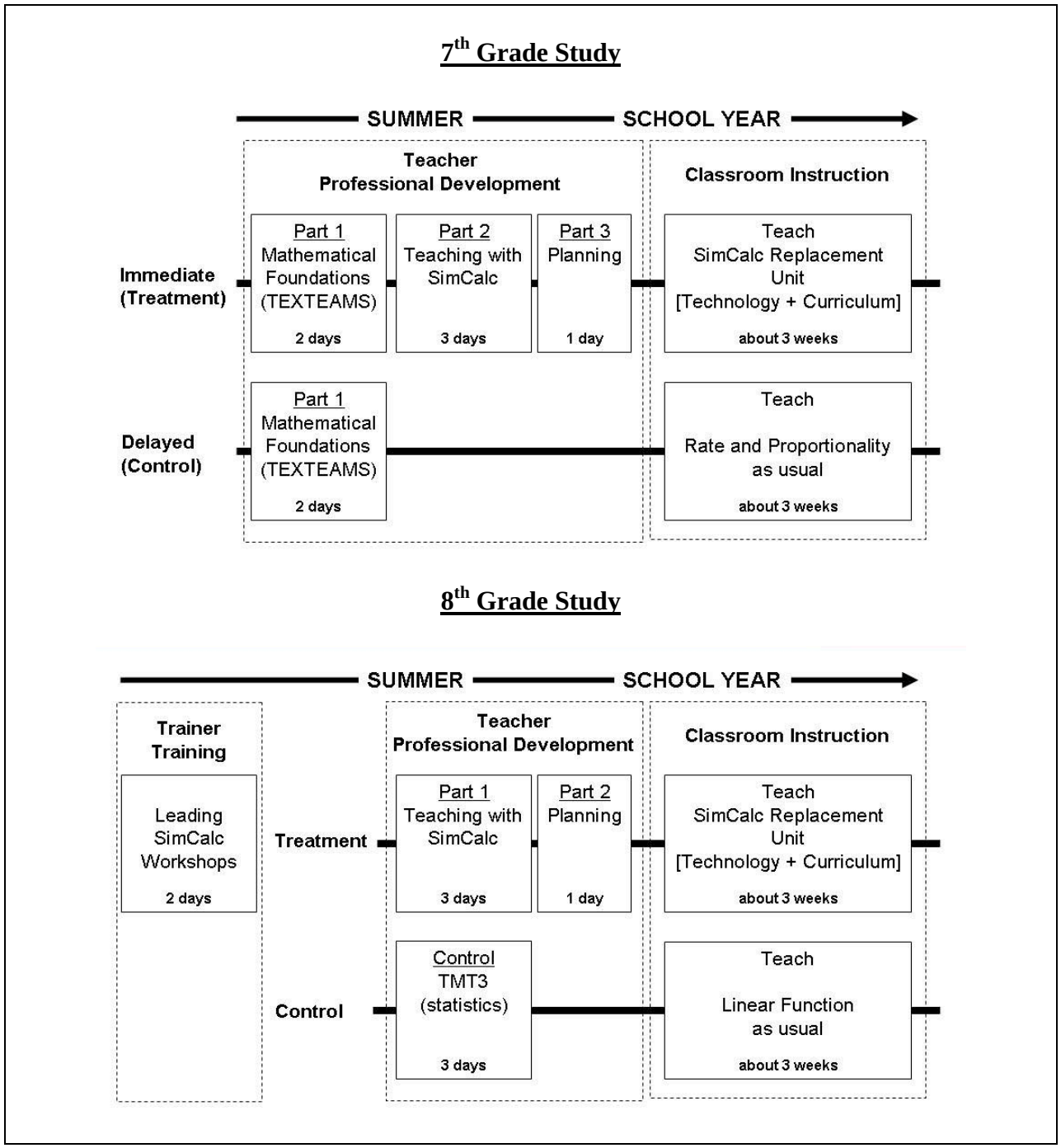
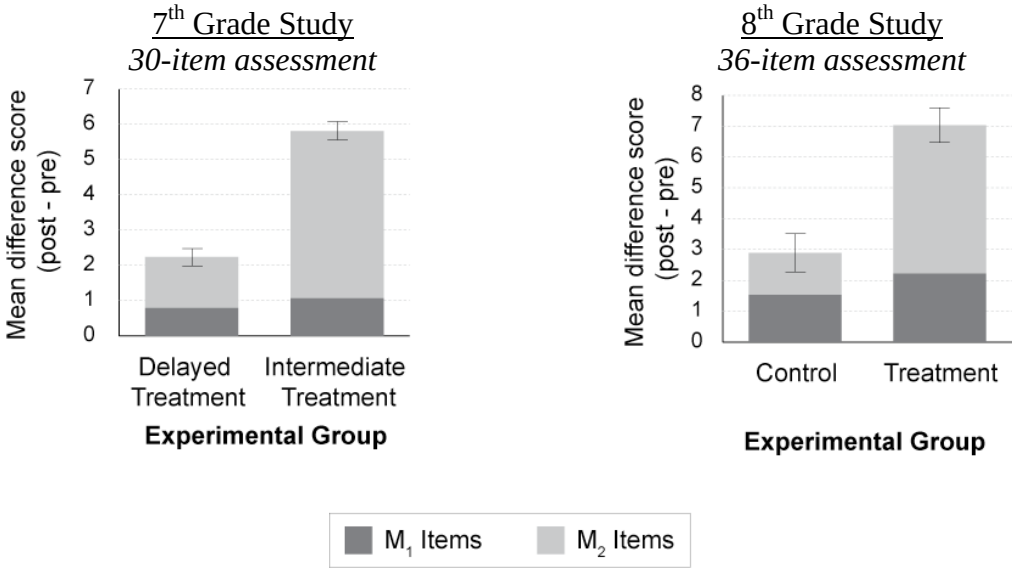


Figure 2. Experimental designs and timelines for the two studies.



	7th Grade Study (N=95)		8th Grade Study (N=56)	
Gain (Posttest-Pretest)	z Statistic	Effect Size	z Statistic	Effect Size
Total Score	9.04*	0.63	5.38*	0.56
M1 Items	1.82	0.10	1.61	0.19
M2 Items	10.03*	0.89	7.62*	0.81

*p<0.0001

Figure 3. Student mean difference scores (± SE of total using HLM) and effect sizes at the student level.